Module 3: Assignment 2

Yanyang Pei (B3)

Introduction

Colorectal cancer (CRC) is one of the most common cancers and is associated with high mortality rates worldwide. Although there's treatment of chemotherapy for patients with late stage CRC and locally advanced rectal cancer, the outcomes are often unsatisfactory. To improve patients' chances of survival, study of new biomarkers have become an urgent need.

MicroRNAs (miRs), small, single-stranded, non-coding RNAs, can regulate mRNA translation and initiate mRNA degradation. Recent research has found abnormal miR levels can be associated with patients with colorectal cancer (CRC) or metastatic colorectal cancer (mCRC). Some miRNAs that have already proven to be biomarkers include miRNA20 (related with later stage cancer), miRNA93, miRNA223 and so on. [1] What's more, some bacterium such as bacteroides fragilis, Escherichia coli and Enterococcus faecalis are studied to act as promising biomarkers for early detection of intestinal cancer. [2]

The aim of the study is to investigate if chosen biomarkers and bacterias are potentially correlated, and determine their relation with different cancer stages. The hypothesis is that bacterial profiles cannot fully substitute for biomarker data, as they may primarily reflect early-stage cancer, while biomarkers are more closely associated with later-stage cancer.

Materials and Methods

Data analyses were done with Python 3.1.2. Dataset containing testing data for 45 bowel cancer patients that are in different stages and have different gender and age. Testing data includes 9 blood biomarkers and 22 bacteria from fecal microbiome sampling.

The data was examined for outliers with a Grubbs's test, and no obvious outliers were found. Empty rows and blanks were removed by python. Cluster Analysis (CA), Shapiro-Wilk test and correlation plot were made to pre-study data property.

PLS plot was done to provide rough insight of the data. The result was compared with PCA analysis and decision boundary. This is followed by building a regression and modeling as well as classification. The model was generated by PLS regression as well as logistic regression, classification ability was studied by plotting confusion matrix and decision boundary plots.

Kruskal-Wallis, ANOVA and univariate regression analysis were performed as extra study to investigate if biomarkers and bacterium are significantly different between the stages.

Results

Preprocess and prestudy

The cluster analysis was performed to study the similarities between various cancer stages, from which it can be seen that stage 1 are are clustered together with stage 2, and stage 3 and

stage4 are also tightly clustered but less spread out, indicating more similarities. Based on this, the later study is focused on grouped stages (early and later).

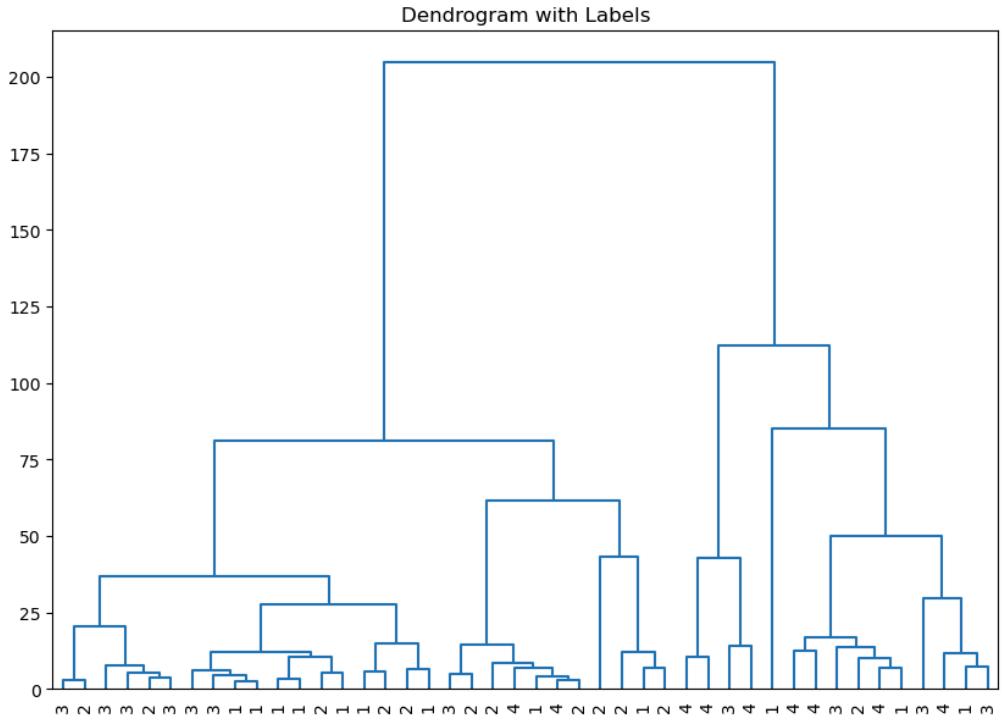

Figure 1. Dendrogram for cancer stage

The normality of biomarkers and bacterium data were studied by Shapiro-Wilk's test, the result is shown in Table A.1 in Appendix. The result shows that almost all of them are not normally distributed, except miRNA15, GDf5, GDf6 in biomarkers and Campylobacter, Blautia, S247, Odoribacter, Clostridiales and Oscillospira in bacterium.

A correlation plot was made to study the relationship within biomarkers and bacterium, the results are shown in Figure A.1 in Appendix.

PLS regression and model study

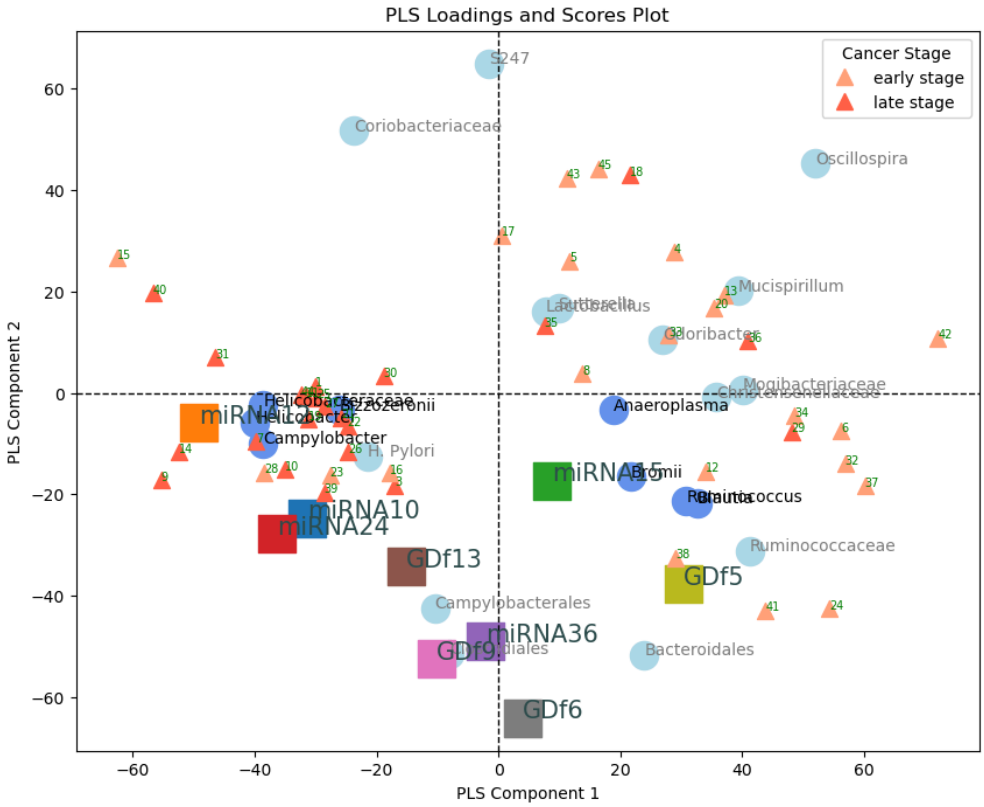

Figure 2. PLS plot showing potential correlation within parameters, where squares represent biomarkers and circles represent bacterium. Bacterium used in effect plots were highlighted.

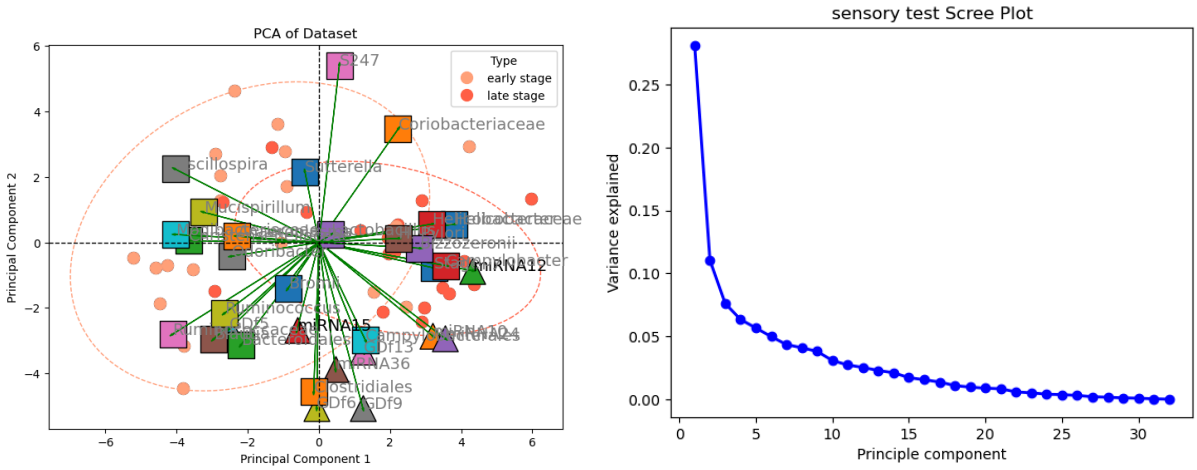

Figure 3. PCA biplot and corresponding scree plot

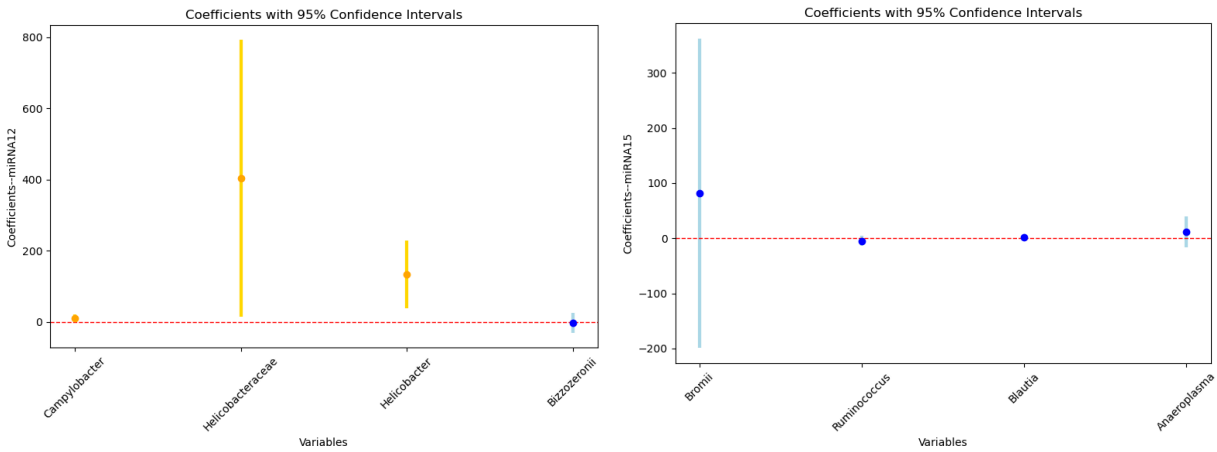

Figure 4. Effect plot for miRNA12 (left) and miRNA15 (right) with their potential correlated bacterium, where yellow color indicates significant relationship.

The PLS plot shown in Figure 2 reveals that late-stage cancer clusters closely with biomarkers such as miRNA12, miRNA10, and miRNA24, whereas early-stage cancer is generally far from most biomarkers, except miRNA15 and GDf5. Overall, it can be seen that the early stage is more correlated with bacteria while the later stage is more correlated with biomarkers.

The PCA biplot shown in Figure 3 reveals a similar correlation that aligns with PLS. It's worth noting that the scree plot shows a very low explained variance for principal components, indicating the system has many dimensions and variables are not too correlated.

The correlation between miRNA12 and surrounding bacteria was examined using effect plots, and the same analysis was applied to miRNA15, the result is shown in Figure 4 and Table 1. The result indicates that most of the bacterium can explain miRNA12, while none of the bacterium are representative of miRNA15.

The decision boundary plot from logistic regression shown in Figure 5 indicates that although miRNA12 and miRNA15 correlate with early and late stages, they might not be predictive in separation of stages since they have a high chance of being misclassified in the model.

Table 1. Correlation and significance summary of biomarkers and bacterium, where red indicates positive correlation.

Stage	Biomarkers	Bacterium	Correlation and Significance
1&2 (early stage)	miRNA12	Campylobacter	*
		Helicobacteraceae	*

		Helicobacter	*
		Bizzozeronii	ns
3&4 (later stage)	miRNA15	Bromii	ns
		Ruminococcus	ns
		Blautia	ns
		Anaeroplasma	ns

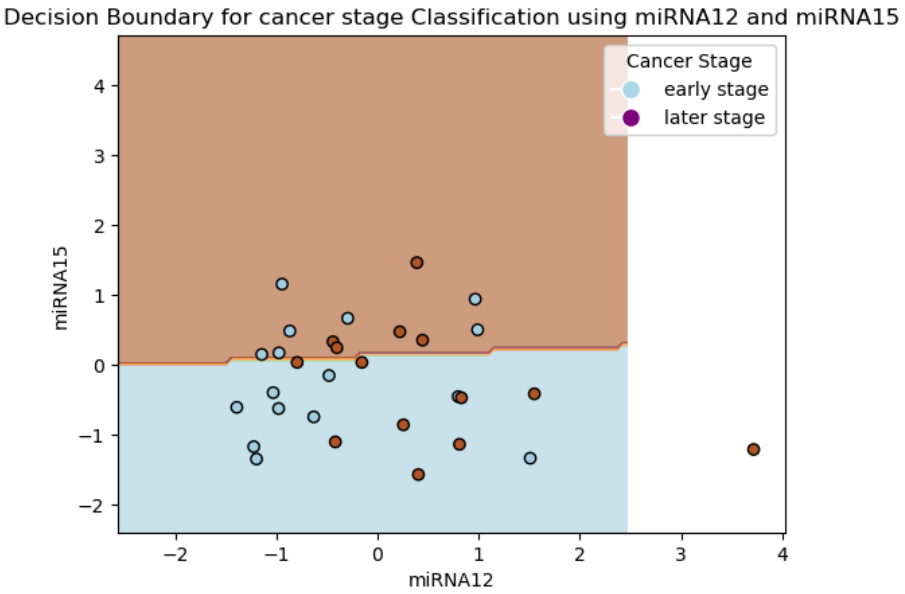

Figure 5. Decision boundary plot for cancer stages and biomarkers.

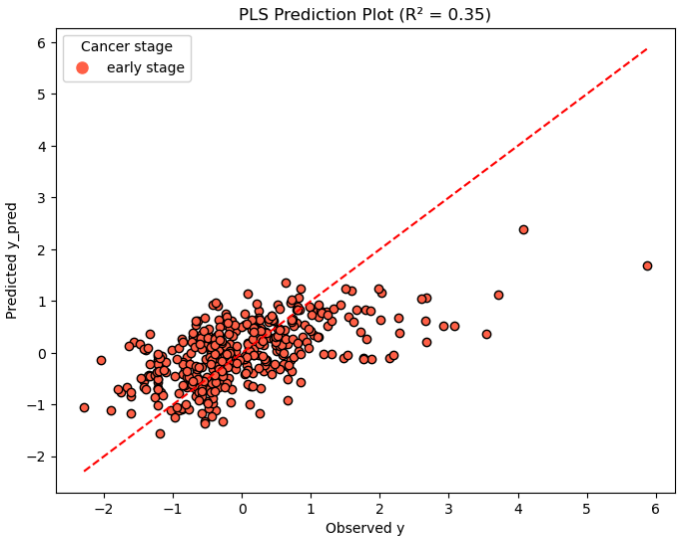

Figure 6. PLS regression of full data.

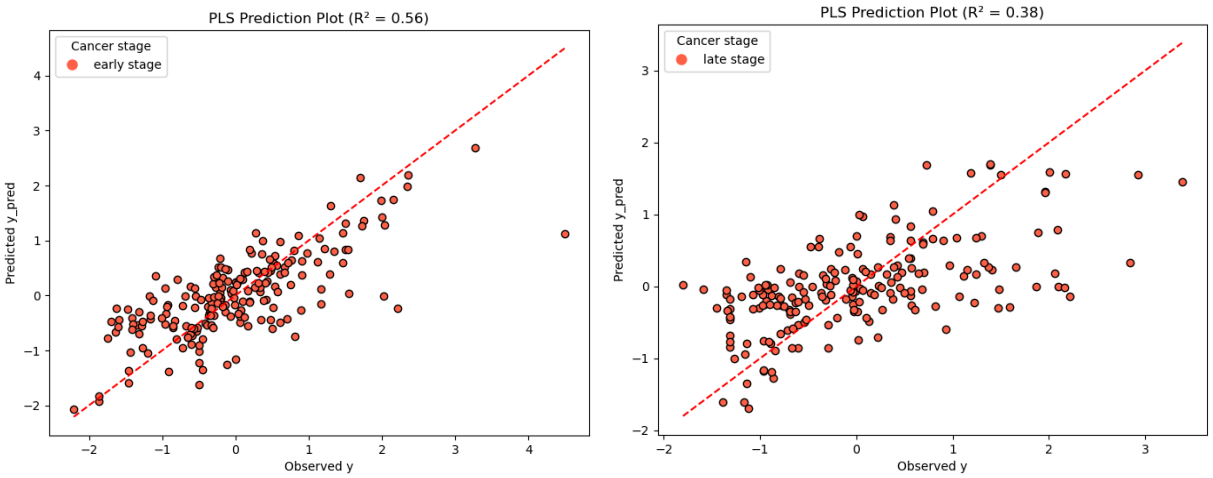

Figure 7. PLS regression plot for early stage (left) and late stage (right) separately

PLS regression was performed both on full data and separate stages, which are shown in Figure 6 and Figure 7. The result revealed that bacterium data cannot represent biomarkers quite well. Since only regression for the early stage gives a decent R2, the cross-validation is only done on it. The results shown in Figure 8, from which it can be seen that the model overfit for PC over 2, but there's an increase after 4, which might just be error caused by more relevant

variation being added.

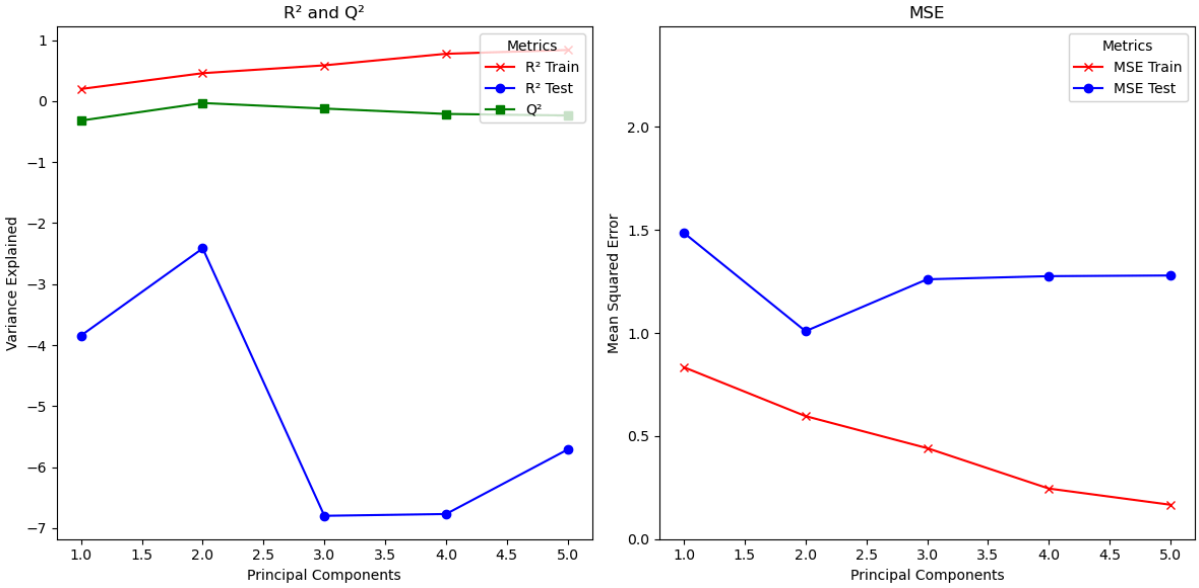

Figure 8. R^2, Q^2 and MSE plot for early stage model

Classification

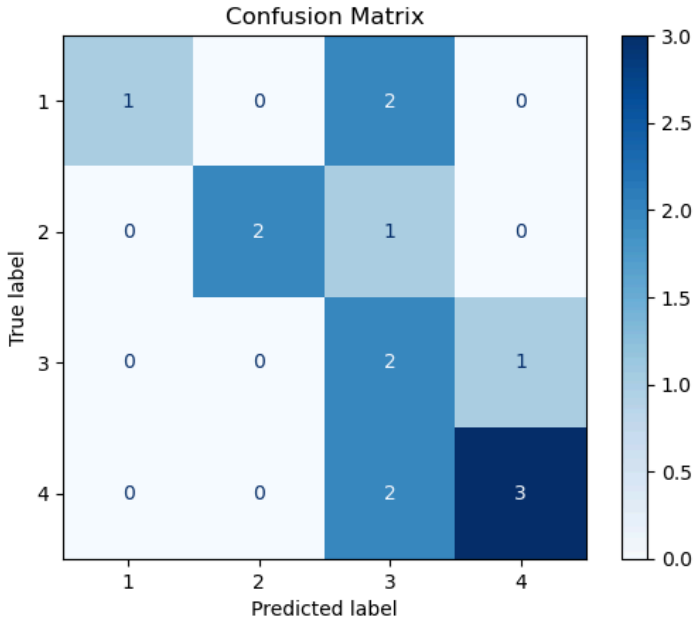

Figure 9. Confusion plot for all cancer stage classified by biomarkers and bacterium

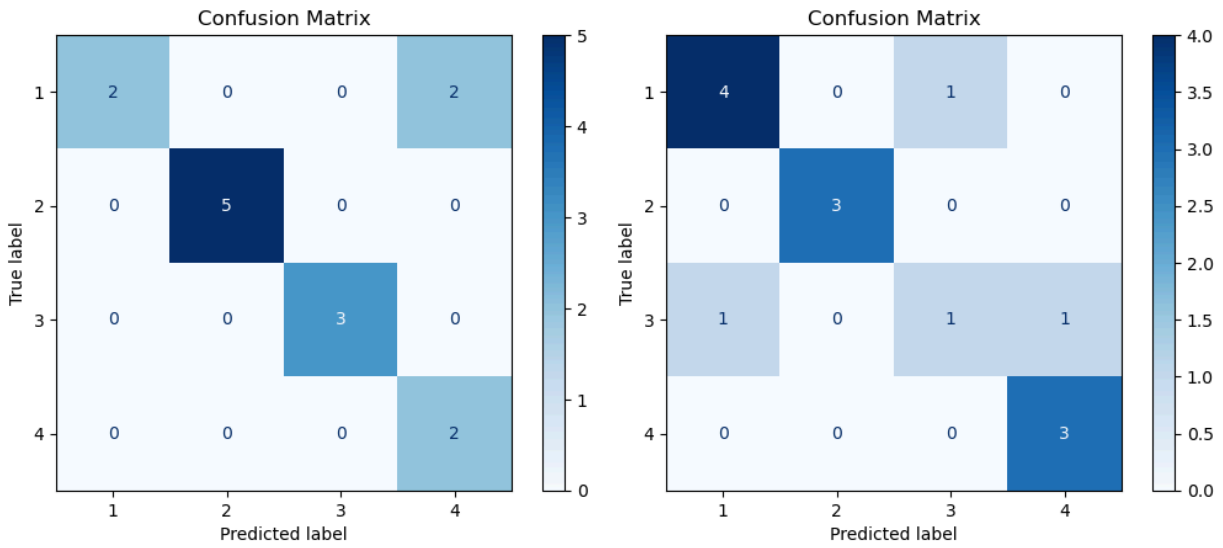

Figure 10. Confusion plot for biomarkers (left) and bacteria (right) separately.

The confusion plots in Figure 9 shows that none of the cancer stages can be completely classified correctly. The accuracy of the model is 0.57 and the precision is 0.76. Confusion plots for biomarkers and bacteria are also made and shown in Figure 10. It's interesting to note that the result aligns with the PLS study, which indicates that biomarkers classify later stage better while bacteria classify early stage better.

Extra study

To further investigate the difference on bacteria and biomarkers for different stages, ANOVA and Kruskal-Wallis tests were performed and bar charts and box plots were made. The results are shown in Table 3 as well as Figure 11-12. It can be concluded that most of the variables identified as significant in the statistical tests (those shown in bar and box plots) align with their position in the PLS plot. It's also noticeable that there's wide error bars in some of the groups, which indicates that variance inside data is large.

Table 3. ANOVA and Kruskal Wallis study result

Bacterium	Significance	Biomarkersme	Significance
Helicobacter	**	miRNA10	**
Coriobacteriaceae	ns	miRNA12	**
Bacteroidales	ns	miRNA15	ns
Campylobacter	*	miRNA24	**
Lactobacillus	ns	miRNA36	ns
Blautia	*	GDf13	*
S247	ns	GDf9	ns
Odoribacter	ns	GDf6	ns
Mucispirillum	**	GDf5	*
Campylobacterales	ns		
Bromii	ns		
Clostridiales	**		
Christensenellaceae	**		
Helicobacteraceae	*		
Bizzozeronii	**		
H. Pylori	*		
Ruminococcaceae	**		
Oscillospira	**		
Ruminococcus	ns		
Mogibacteriaceae	**		
Sutterella	ns		
Anaeroplasma	**		

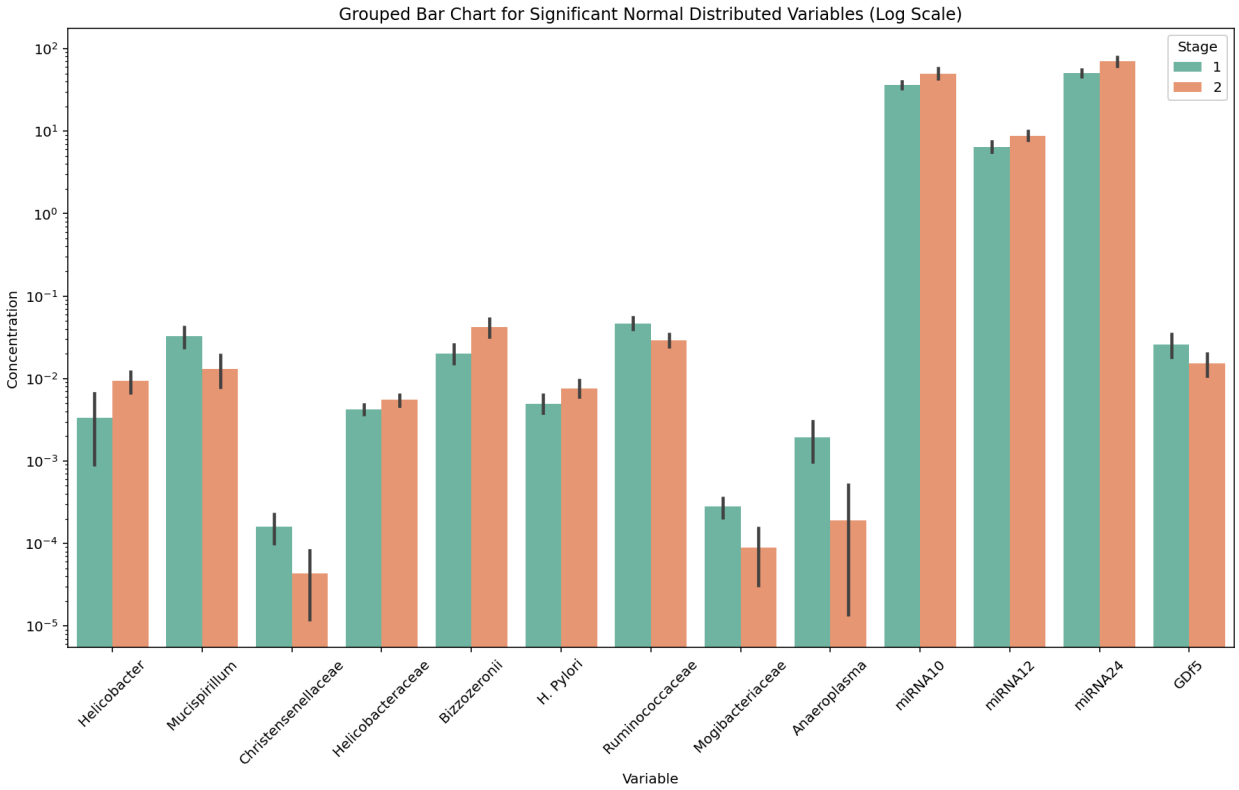

Figure 11. Barchart for normal distributed bacterium and biomarkers, error bar plotting 95% confidence interval

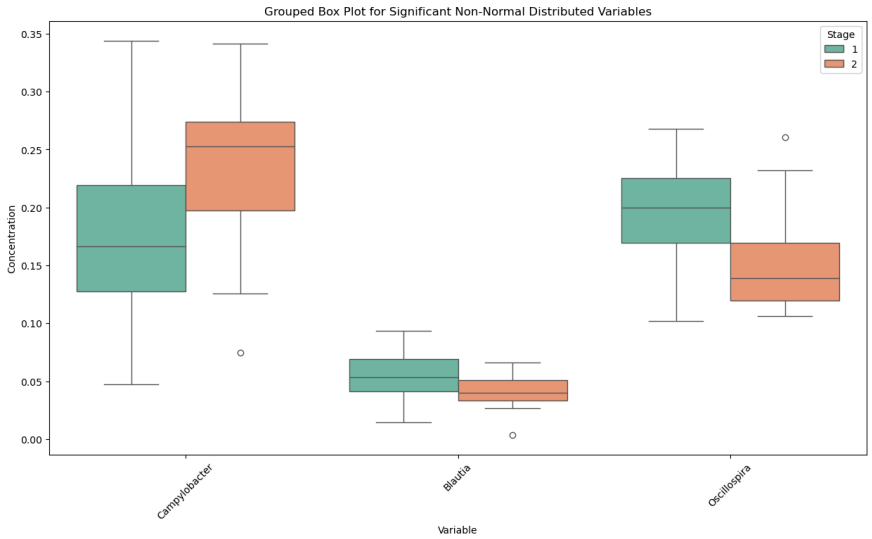

Figure 12. Box chart for non-normal distributed bacterium and biomarkers

Discussion

The study shows that bacteria sampling cannot fully replace biomarkers for identifying different stages of bowel cancer well, but some of them have high relevance. So using a model to predict biomarkers results from bacteria behavior is not very effective. From PLS, PCA and also model classification it can be seen that bacteria classifies early stage cancer better while biomarkers classifies later stage cancer better. This differentiation supports the hypothesis. Similar results as previous studies, stating that bacteria are potential indicators for intestine disease [3], and biomarkers such as miRNAs as the predictor of cancer progression [4].

Since the samples were taken from patients from different ages and gender, there might be an impact on the data grouping. If more time and energy is available, study would continue on separating gender and age groups to see if the model works well on any of them.

Further study can be deep down to investigate the mechanisms of these markers. And even more, studying if diet has a significant impact on the amount of bacteria in the colonel can be a very interesting and insightful topic.

Conclusion

This study examined the potential of biomarkers and fecal bacteria to distinguish between different stages of bowel cancer. The results indicate that while bacterial profiles are more associated with early-stage cancer, blood biomarkers demonstrate a stronger correlation with later stages. This finding suggests that bacteria alone cannot fully replace biomarkers in cancer staging but could complement biomarkers to improve diagnostic accuracy. However, there’s still limitations in the study, which need to be further studied.

Appendix

	type	p_value	normality
0	miRNA10	0.000147	Not Normally Distributed
1	miRNA12	0.002447	Not Normally Distributed
2	miRNA15	0.624177	Normally Distributed
3	miRNA24	0.00057	Not Normally Distributed
4	miRNA36	4.18E-11	Not Normally Distributed
5	GDf13	0.003445	Not Normally Distributed
6	GDf9	0.876538	Normally Distributed
7	GDf6	0.452958	Normally Distributed
8	GDf5	0.000189	Not Normally Distributed
9	Helicobact	2.94E-06	Not Normally Distributed
10	Coriobacte	0.000326	Not Normally Distributed
11	Bacteroida	0.000886	Not Normally Distributed
12	Campyloba	0.477465	Normally Distributed
13	Lactobacil	1.36E-14	Not Normally Distributed
14	Blautia	0.946431	Normally Distributed
15	S247	0.069447	Normally Distributed
16	Odoribacte	0.499629	Normally Distributed
17	Mucispirill	0.000635	Not Normally Distributed
18	Campyloba	0.000394	Not Normally Distributed
19	Bromii	0.005533	Not Normally Distributed
20	Clostridiale	0.900511	Normally Distributed
21	Christense	5.86E-07	Not Normally Distributed
22	Helicobact	0.024644	Not Normally Distributed
23	Bizzozeror	6.59E-06	Not Normally Distributed
24	H. Pylori	4.69E-05	Not Normally Distributed
25	Ruminoco	0.000803	Not Normally Distributed
26	Oscillospir	0.067986	Normally Distributed
27	Ruminoco	0.03066	Not Normally Distributed
28	Mogibacte	9.87E-05	Not Normally Distributed
29	Sutterella	1.74E-14	Not Normally Distributed
30	Anaeropla	5.90E-10	Not Normally Distributed

Table A.1. Normality test result of data from Shapiro-Wilk

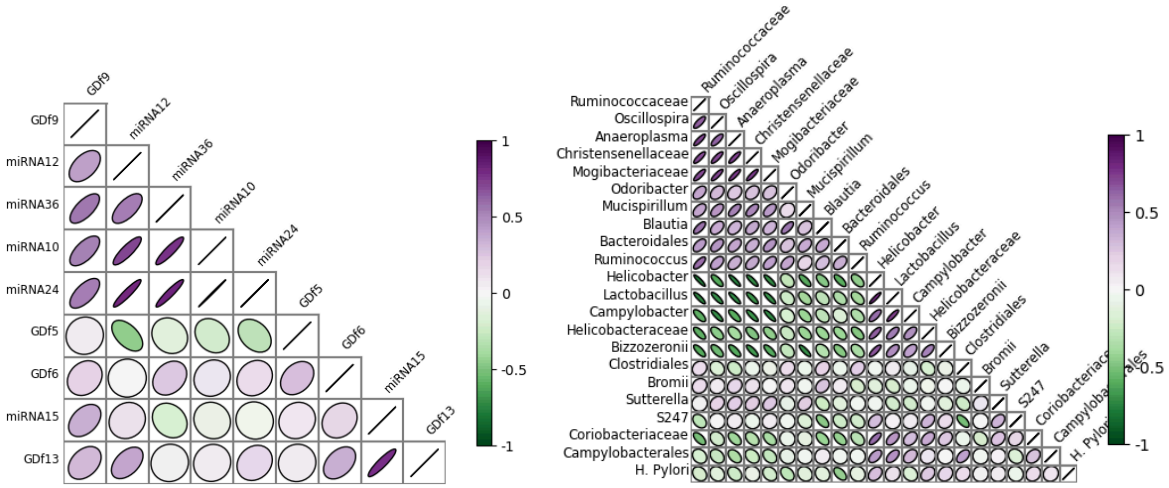

Figure A.1. Correlation plot for biomarkers and bacterium

Reference

[1] Yang, I. P., Yip, K. L., Chang, Y. T., Chen, Y. C., Huang, C. W., Tsai, H. L., Yeh, Y. S., & Wang, J. Y. (2023). MicroRNAs as Predictive Biomarkers in Patients with Colorectal Cancer Receiving Chemotherapy or Chemoradiotherapy: A Narrative Literature Review. Cancers, 15(5), 1358. https://doi.org/10.3390/cancers15051358
[2] Wong, S. H., & Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nature reviews. Gastroenterology & hepatology*, *16*(11), 690–704. https://doi.org/10.1038/s41575-019-0209-8
[3] Alhinai, E. A., Walton, G. E., & Commane, D. M. (2019). The Role of the Gut Microbiota in Colorectal Cancer Causation. *International journal of molecular sciences*, *20*(21), 5295. https://doi.org/10.3390/ijms20215295
[4] Schetter, A. J., Okayama, H., & Harris, C. C. (2012). The role of microRNAs in colorectal cancer. Cancer journal (Sudbury, Mass.), 18(3), 244–252. https://doi.org/10.1097/PPO.0b013e318258b78f

Acknowledgements

AI was used in code writing.